Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

BSC-HIRLAM collaboration: HARMONIE code profiling roadmap and current status

Xavier Yepes-Arbós
Mario C. Acosta
Kim Serradell

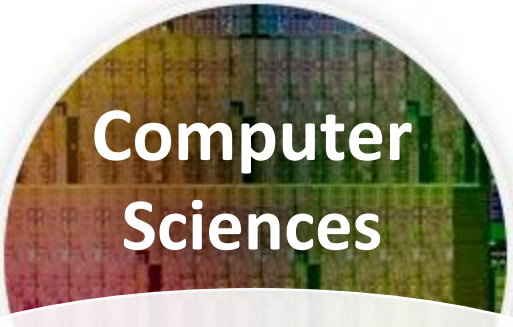
18/06/2019

HARMONIE System Working Week, ECMWF, UK

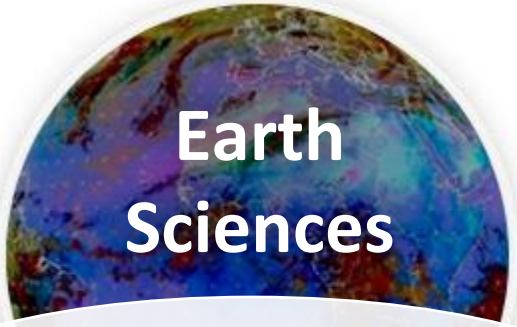
Who we are

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Mission of BSC Scientific Departments



To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency



To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications

To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)

To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)

Earth Sciences

Environmental modelling and forecasting, with a particular focus on weather, climate and air quality

Modeling **air quality** and **Sand and Dust Storms** Processes from urban to Global and the impacts on weather, health and ecosystems

Research and Forecasts

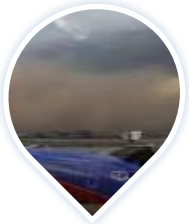
Climate predictions system from subseasonal-to-decadal forecasts

Service Users Sectors


Infrastructures


Solar Energy

Urban development


Transport

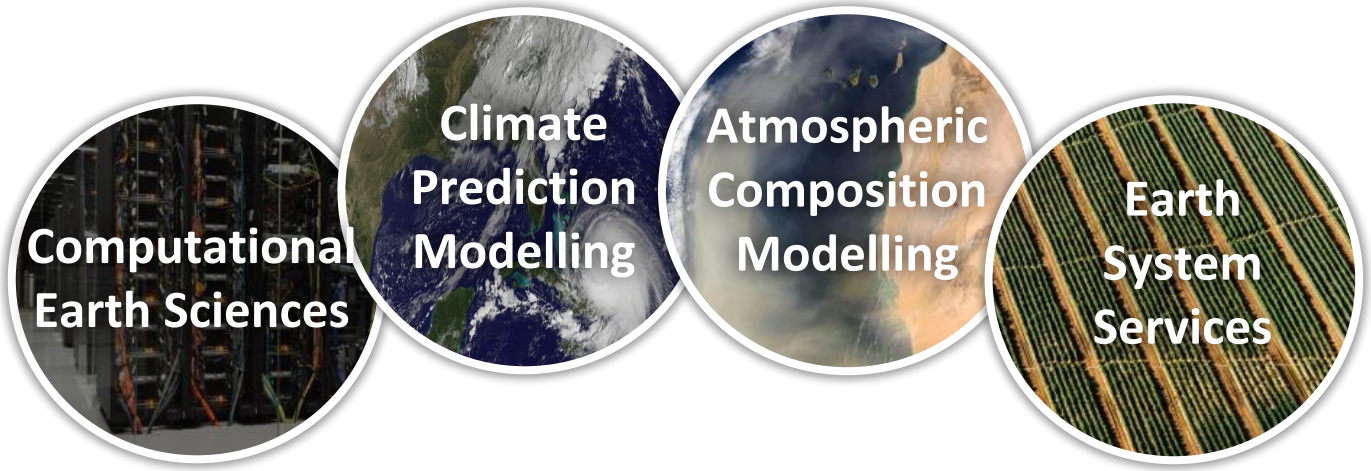

Wind Energy

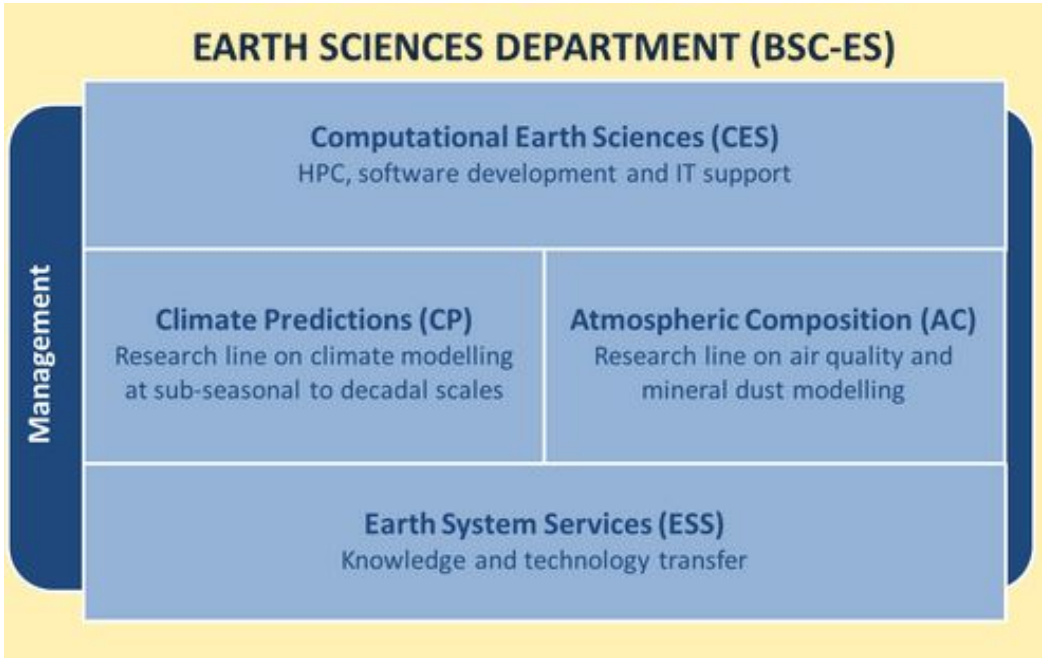

Agriculture


Insurance

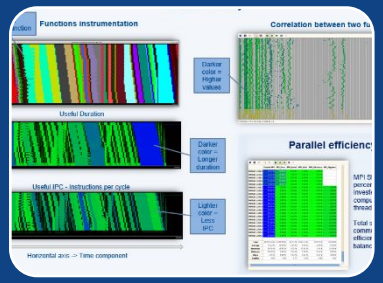
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Earth Sciences Groups



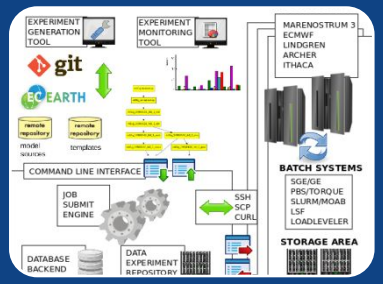


Computational Earth Sciences Group



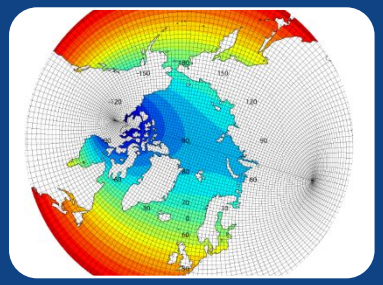
Performance Team

- Provide HPC Services (profiling, code audit, …)
- Apply new computational methods



Models and Workflows Team

- Development of HPC user-friendly software framework
- Support the development of atmospheric research software



Data and Diagnostics Team

- Big Data in Earth Sciences
- Provision of data services
- Visualization

Roadmap

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-HIRLAM collaboration

- The BSC and the HIRLAM consortium signed a contract for a 1 year project to perform a complete code profiling of the HARMONIE-AROME model and the Data Assimilation system, starting on June 2019
- The project consists of two phases:
  - 1st: Basic profiling analysis
  - 2nd: Perform a complete analysis according to the results from the first phase

Scope of the phase 1

- Duration: 4 months
- Prepare selected configurations to be deployed with Extrae on cca/ccb at ECMWF, a Cray XC40 machine
- Perform a basic analysis of the HARMONIE-AROME Forecast model and the Data assimilation execution
  - Use different computational metrics: IPC, useful duration, MPI overhead, cache misses, etc
  - Identify the different parts of the trace with regard the code being executed
- Deliver a complete document with the results and feedback to decide the main goals for the profiling analysis of the phase 2

Scope of the phase 2

- Duration: 8 months after completion of phase 1
- Complete profiling analysis according to the results obtained from phase 1
- Training:
  - Prepare basic tutorials based on coarser HARMONIE configurations
  - Prepare a physical event to perform a training for the users
- Prepare complete documentation:
  - Online follow-up meetings to detect deviations and correct if needed
  - Write a final document describing all the tasks carried on
- Presence in the HIRLAM System group meetings: dissemination and feedback

Code profiling methodology

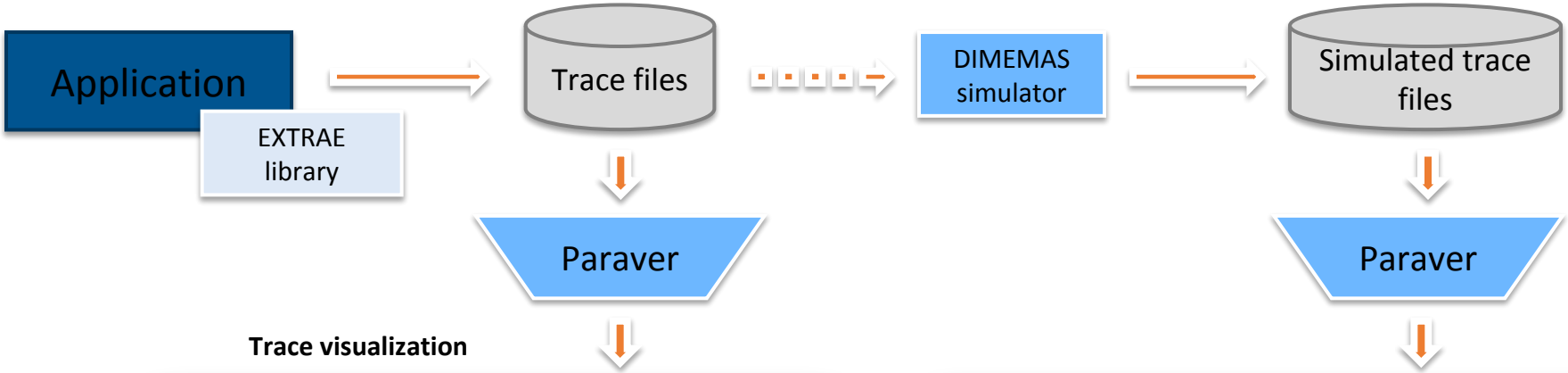
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Profiling methodology overview

- Scalability tests: MPI, OpenMP, Tiling
- Evaluate deployment efficiency: compilation flags
- Affinity tests: find a proper placement for MPI processes
- Profile analysis: user functions calls, statistics…
- Trace analysis: MPI, hardware counters, communication…
- Performance simulation: evaluate the code under machine changes
- Validation tests: check code correctness if optimized

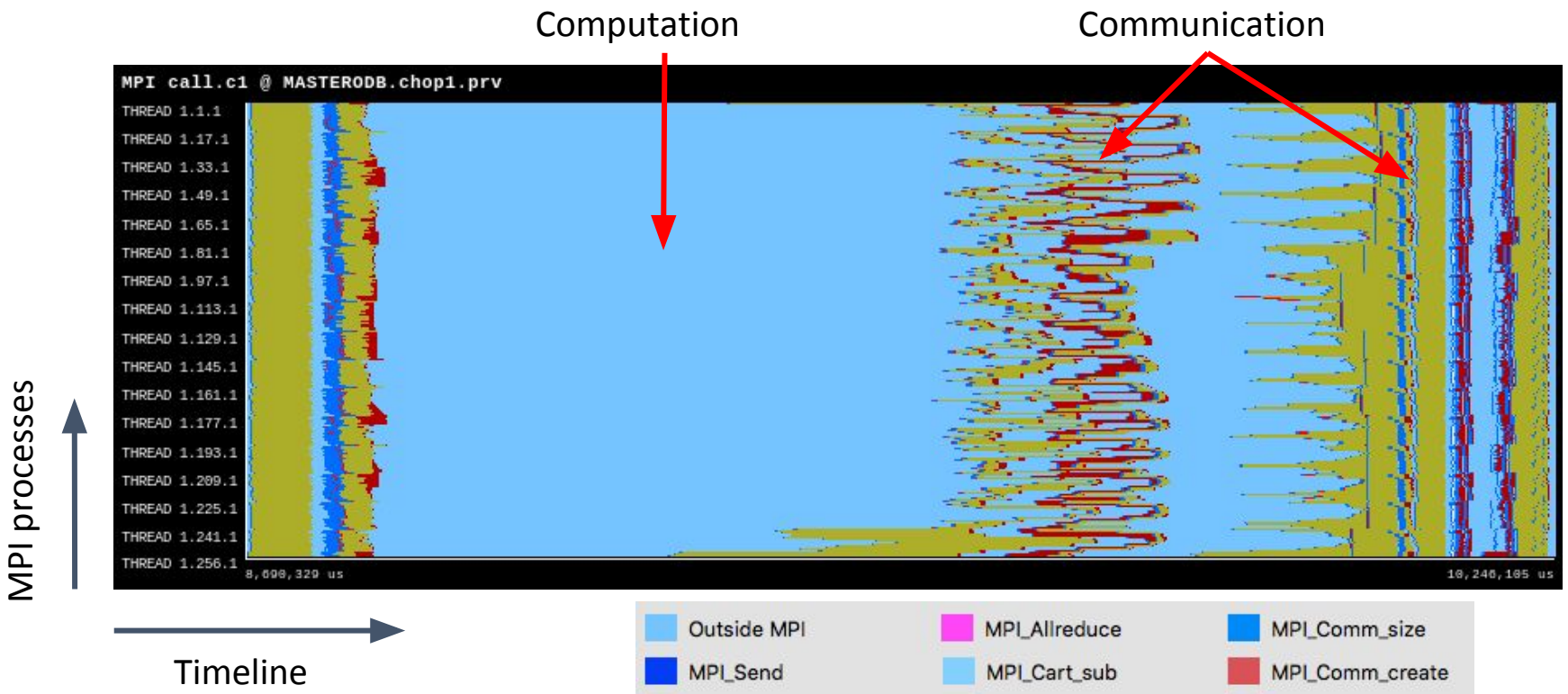
BSC performance tools

- Since 1991
- Based on traces
- Open Source: http://www.bsc.es/paraver
- **Extrae**: Package that generates Paraver trace-files for a post-mortem analysis
- **Paraver**: Trace visualization and analysis browser
  - Includes trace manipulation: Filter, cut traces
- **Dimemas**: Message passing simulator



DIMEMAS generated trace. Target = ideal machine

How a trace looks like: basic overview



Current status

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Deployment

- Git develop branch
- The first deployed configuration is the default one
- DKCOEXP domain (2.5km and 65 vertical levels)
- Run on cca at ECMWF
- GNU compiler
- Four different scenarios have been tested by combining OpenMP and the I/O server
- Only one of them worked with Extrae

1st scenario

- Default setup: 256 MPI (16x16) and no OpenMP. 2 IO servers
- It only crashes when it is used with Extrae
- The I/O server voluntarily aborts the execution
- Error:

```
ABORT!    1 IO_SERV_GET_REQ: SIZE MISMATCH
ABOR1 CALLED
IO_SERV_GET_REQ: SIZE MISMATCH
MPL_ABORT: CALLED FROM PROCESSOR      1 THRD     1
MPL_ABORT: THRD           1   IO_SERV_GET_REQ: SIZE MISMATCH
[myproc#1,tid#1,pid#53203]: 1081 MB (maxheap), 276 MB
(maxrss), 0 MB (maxstack), walltime = 1554295600.04s
[myproc#1,tid#1,pid#53203]:  IO_SERV_RUN
[myproc#1,tid#1,pid#53203]:   IO_SERV_RUN_MF
[myproc#1,tid#1,pid#53203]:    IO_SERV_RECV_REQ
[myproc#1,tid#1,pid#53203]:     IO_SERV_SYNC_SORT
[myproc#1,tid#1,pid#53203]:      IO_SERV_GET_REQ
```

2nd scenario

- 256 MPI (16x16) and no OpenMP. No IO servers
- It worked with Extrae and a trace was generated

3rd scenario

- 100 MPI (10x10) and 4 OpenMP = 400. No IO servers. It does not use FLAKE
- It is not stable. There were run 3 instances, where 2 of them crashed and 1 properly finished
- First error:

```
Rank 95 [Wed Apr  3 15:42:56 2019] [c2-1c2s9n0] Fatal error in MPI_Recv: Other MPI error, error stack:
MPI_Recv(212).................: MPI_Recv(buf=0x2aacca7995d0, count=4200, dtype=0x4c000829, src=1,
tag=25002, comm=0x84000004, status=0x7fffff78d2d0) failed
MPIDI_CH3I_Progress(537).......:
MPID_nem_mpich_test_recv(989)..:
MPID_nem_gni_poll(1554)........:
MPID_nem_gni_check_recvCQ(1448):
MPID_nem_gni_process_recv(1298): GNI_SmsgGetNextWTag (GNI_RC_INVALID_PARAM)
aborting job:
Fatal error in MPI_Recv: Other MPI error, error stack:
MPI_Recv(212).................: MPI_Recv(buf=0x2aacca7995d0, count=4200, dtype=0x4c000829, src=1,
tag=25002, comm=0x84000004, status=0x7fffff78d2d0) failed
MPIDI_CH3I_Progress(537).......:
MPID_nem_mpich_test_recv(989)..:
MPID_nem_gni_poll(1554)........:
MPID_nem_gni_check_recvCQ(1448):
MPID_nem_gni_process_recv(1298): GNI_SmsgGetNextWTag (GNI_RC_INVALID_PARAM)
[NID 02084] 2019-04-03 15:42:57 Apid 334084920: initiated application termination
```

3rd scenario (2)

- Second error:

```
***Received signal = 11 and ActivatED SIGALRM=14 and calling alarm(10), time =1553846291.95
[myproc#29,tid#1,pid#65932,signal#11(SIGSEGV)]: Received signal :: 8439MB (heap), 1285MB (rss),
0MB (stack), 0 (paging), nsigs 1, time 1553846291.95
tid#1 starting drhook traceback, time =1553846291.95
[myproc#29,tid#1,pid#65932]: 8439 MB (maxheap), 1285 MB (maxrss), 0 MB (maxstack), walltime =
1553846291.95s
[myproc#29,tid#1,pid#65932]:  MASTER
[myproc#29,tid#1,pid#65932]:   CNT0
[myproc#29,tid#1,pid#65932]:    CNT1
[myproc#29,tid#1,pid#65932]:     CNT2
[myproc#29,tid#1,pid#65932]:      CNT3
[myproc#29,tid#1,pid#65932]:       CNT4
[myproc#29,tid#1,pid#65932]:        STEPO
[myproc#29,tid#1,pid#65932]:         SCAN2M
[myproc#29,tid#1,pid#65932]:          GP_MODEL_STACK
[myproc#29,tid#1,pid#65932]:           GP_MODEL
[myproc#29,tid#1,pid#65932]:            CALL_SL_STACK
[myproc#29,tid#1,pid#65932]:             CALL_SL
[myproc#29,tid#1,pid#65932]:              SLCOMM
[myproc#29,tid#1,pid#65932]:               SLCOMM:SLCOMM_INT
[gdb__sigdump] : Received signal#11(SIGSEGV), pid=65932 , it=1 : sigcontextptr=0x7fffff78aa00
```

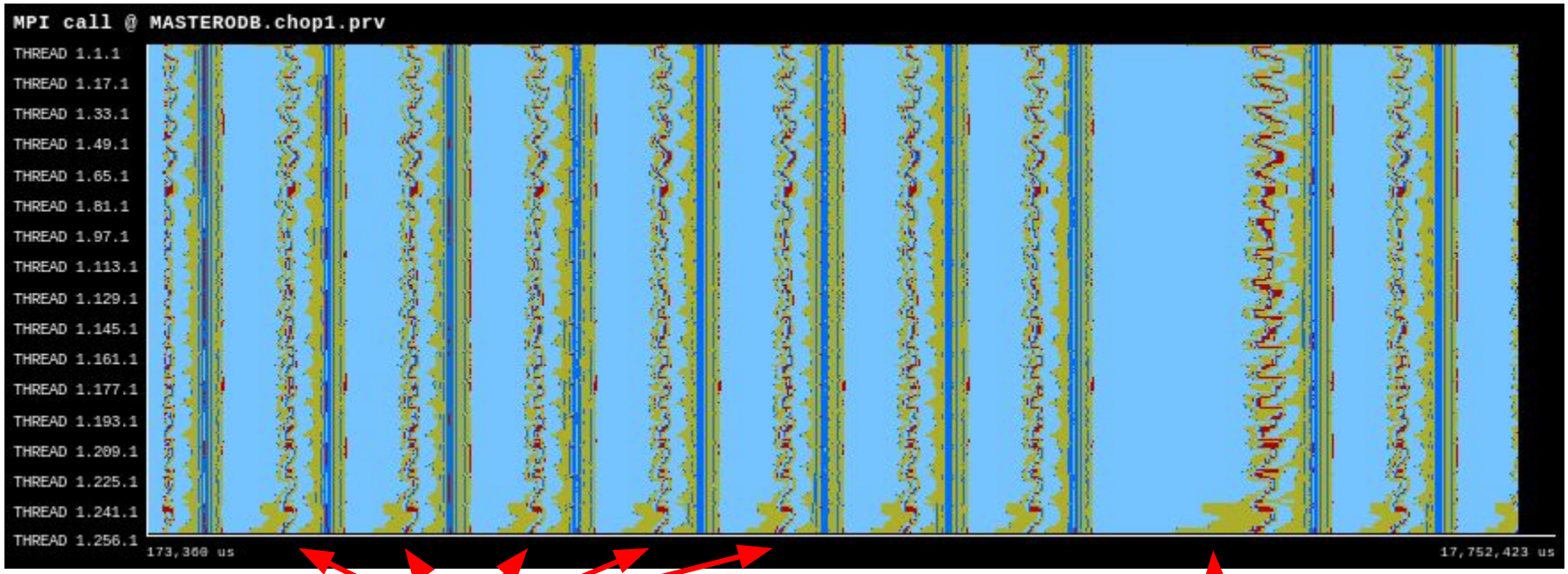
4th scenario

- 100 MPI (10x10) and 4 OpenMP = 400. 2 IO servers. It does not use FLAKE
- As in the 1st scenario, it only crashes when it is used with Extrae
- The I/O server voluntarily aborts the execution
- Error:

```
ABORT!    1 IO_SERV_GET_REQ: SIZE MISMATCH
ABOR1 CALLED
IO_SERV_GET_REQ: SIZE MISMATCH
MPL_ABORT: CALLED FROM PROCESSOR      1 THRD     1
MPL_ABORT: THRD           1   IO_SERV_GET_REQ: SIZE MISMATCH
[myproc#1,tid#1,pid#53203]: 1081 MB (maxheap), 276 MB
(maxrss), 0 MB (maxstack), walltime = 1554295600.04s
[myproc#1,tid#1,pid#53203]:  IO_SERV_RUN
[myproc#1,tid#1,pid#53203]:   IO_SERV_RUN_MF
[myproc#1,tid#1,pid#53203]:    IO_SERV_RECV_REQ
[myproc#1,tid#1,pid#53203]:     IO_SERV_SYNC_SORT
[myproc#1,tid#1,pid#53203]:      IO_SERV_GET_REQ
```

HARMONIE general overview

- Trace from the 2nd scenario
- Several time steps were traced

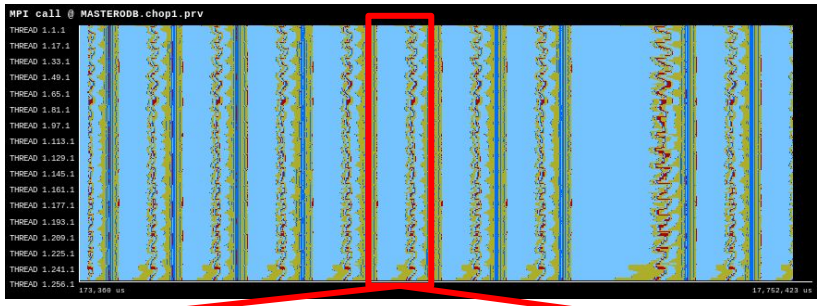


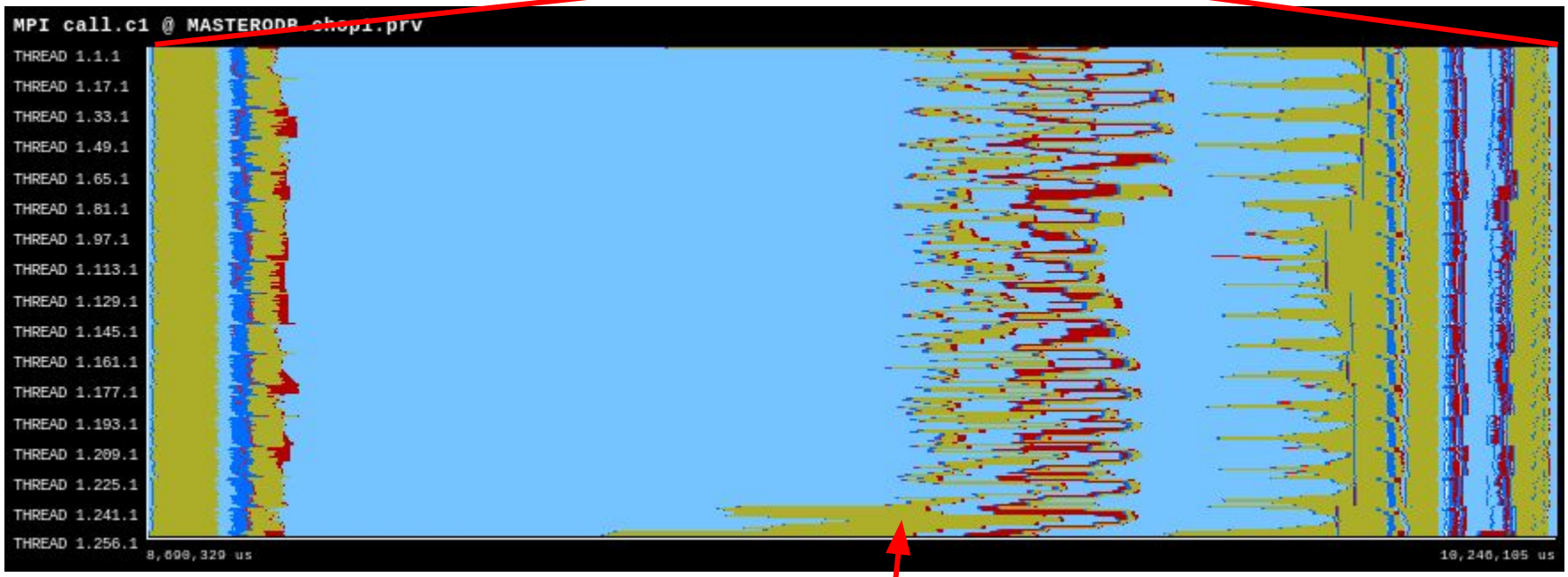
Regular time steps

Larger time step (it is probably computing radiation)

HARMONIE MPI calls

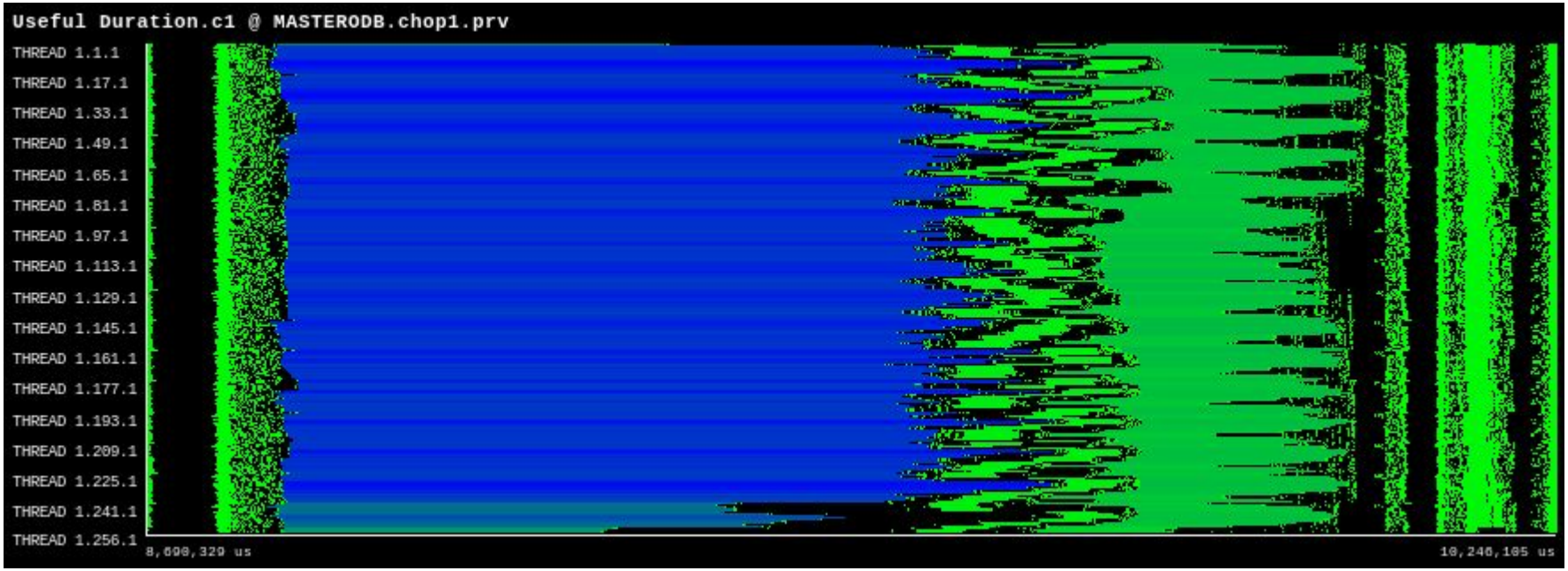
1 regular time step





Workload imbalance

HARMONIE useful duration

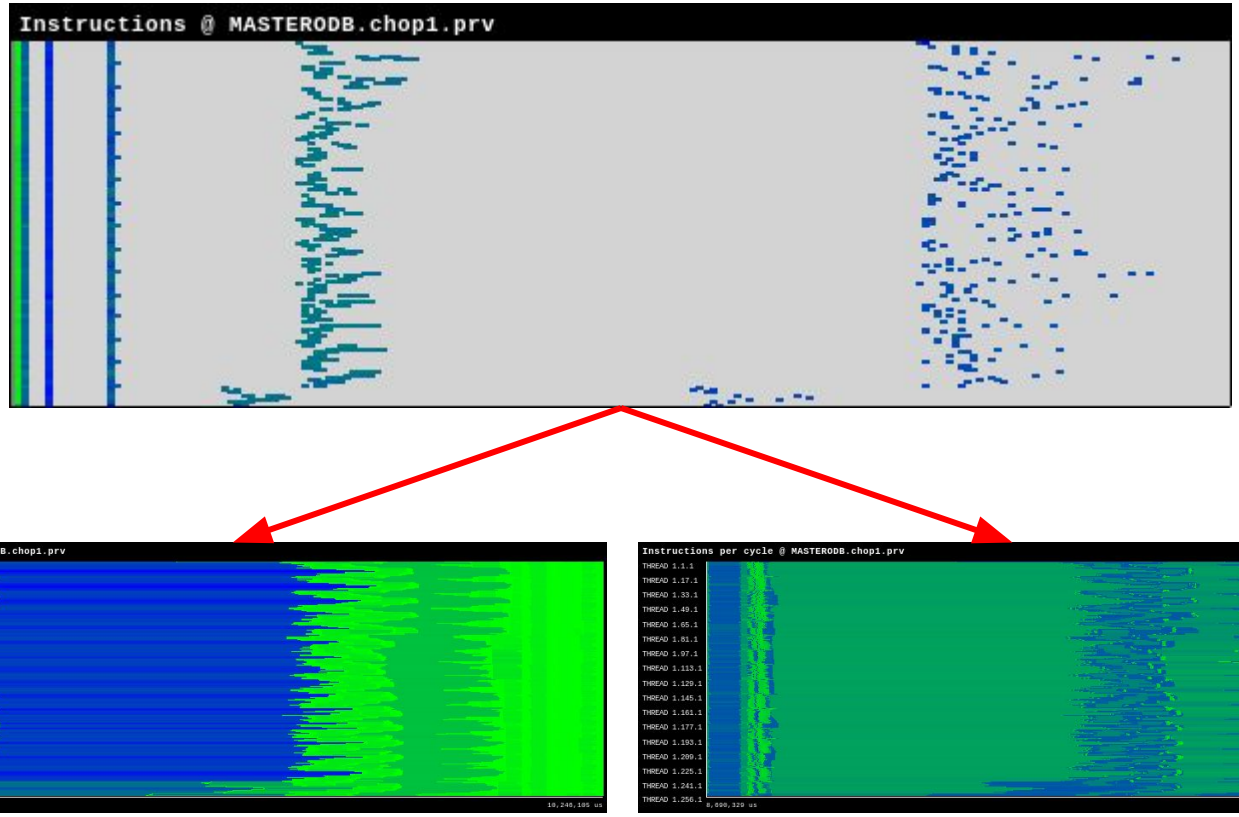


Shorter computation burst

Larger computation burst

HARMONIE Inst.-IPC histogram

Instructions and IPC correlation histogram

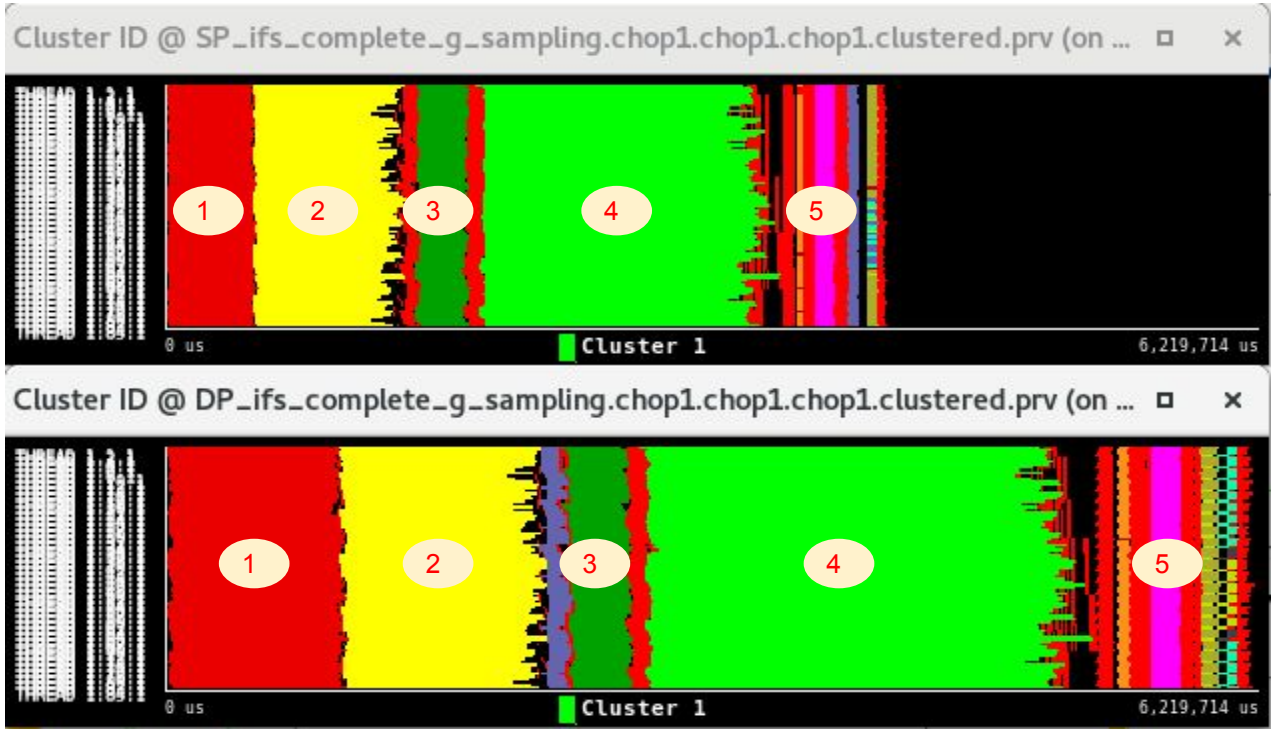


Number of instructions

Instructions per cycle (IPC)

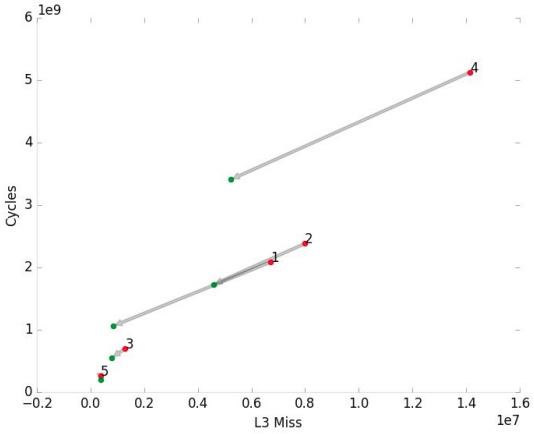
Other BSC tools: Clustering

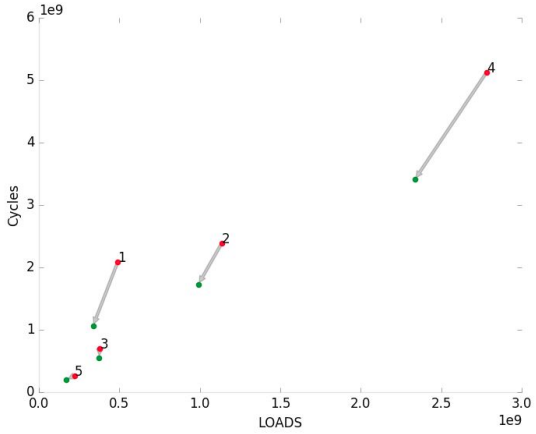
Single precision vs double precision in IFS

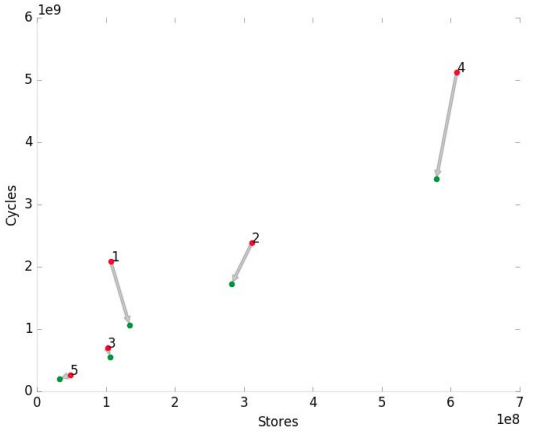


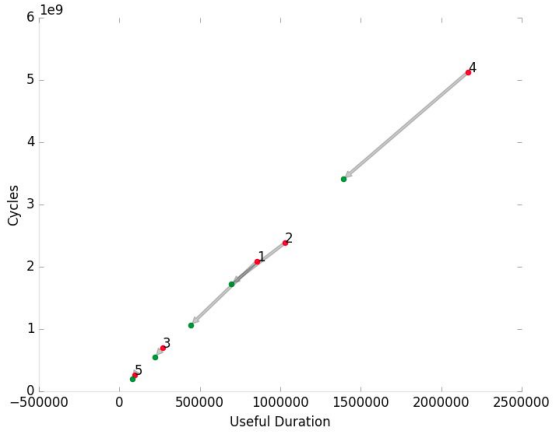
Other BSC tools: Tracking

Single precision vs double precision in IFS



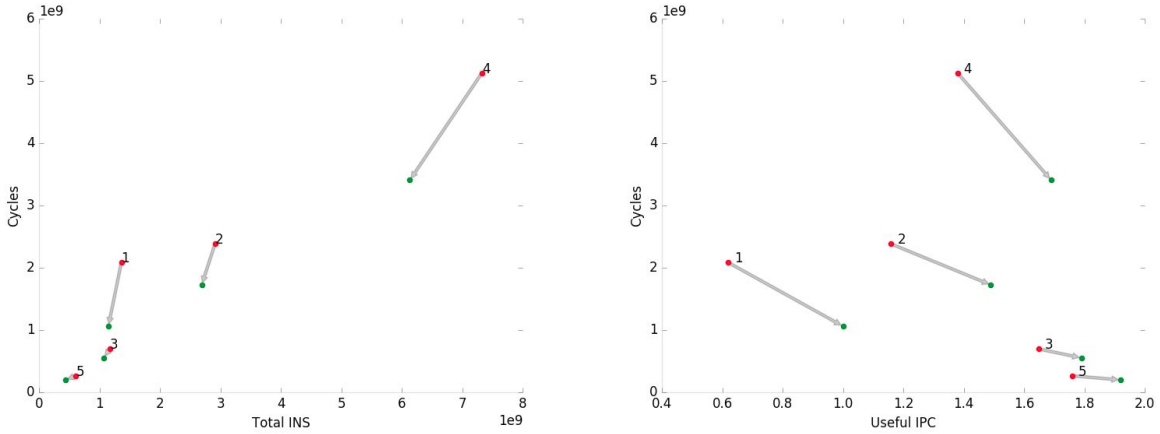






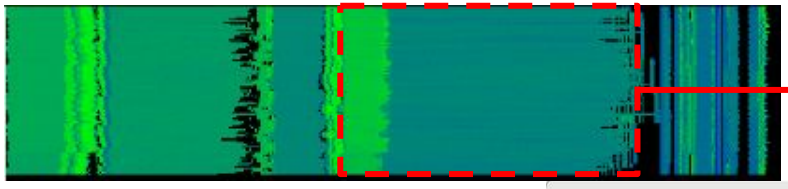
Other BSC tools: Tracking

Single precision vs double precision in IFS



Other BSC tools: Folding

Single precision vs double precision in IFS



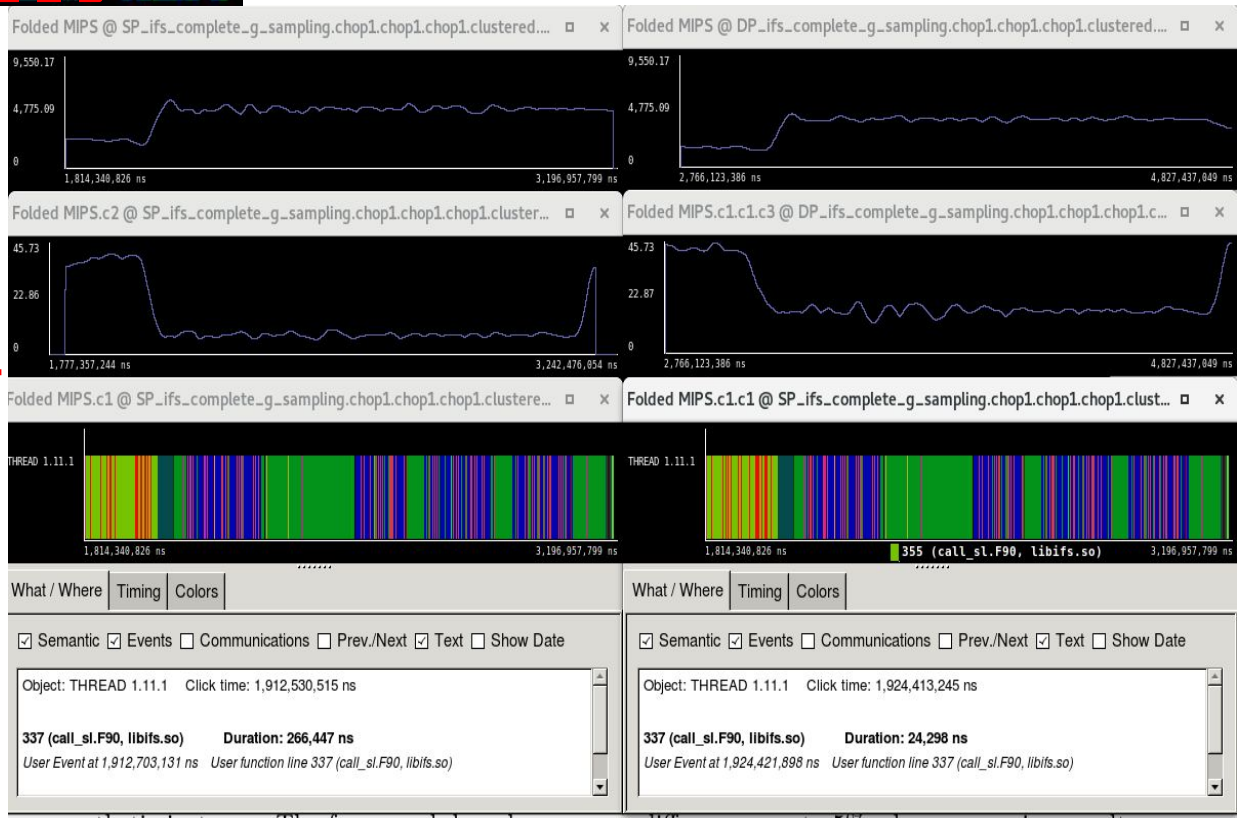
Region studied using sampling+folding

TOT_INS

TOT_CACHE_MISSES

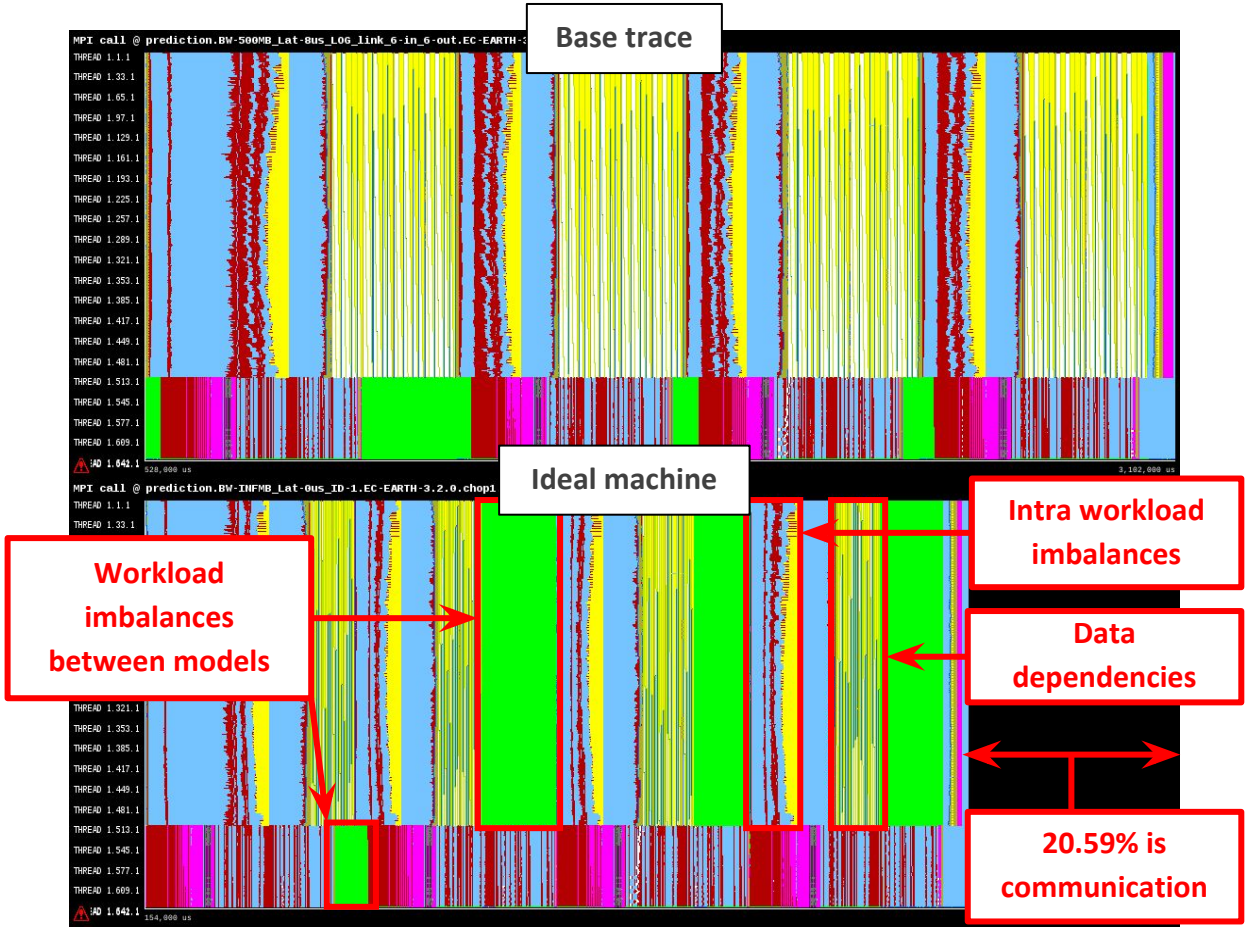
Connection to the code

USER_FUNCTION_LINE



Other BSC tools: Dimemas

Ideal machine test (infinite bandwidth, no latency) in EC-Earth



Discussion

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

List of things to discuss

- To summarize, we ran the default configuration (DKCOEXP) with these 4 variants or scenarios:
  - 1st: 256 MPI (16x16) and no OpenMP. 2 IO servers. Crashes w/ Extrae
  - 2nd: 256 MPI (16x16) and no OpenMP. No IO servers. Works
  - 3rd: 100 MPI (10x10) and 4 OpenMP = 400. No IO servers. It does not use FLAKE. Unstable
  - 4th: 100 MPI (10x10) and 4 OpenMP = 400. 2 IO servers. It does not use FLAKE. Crashes w/ Extrae
- It is very important to have a working and stable configuration to apply all the techniques of the profiling methodology in order to perform a successful study for both phases 1 and 2
- Stable branch to profile?
- Configuration to be used -> DKCOEXP, METCOOP25C… ?

List of things to discuss (2)

- Compiler to be used -> Intel, gcc?
- Generic issues -> systemcore-bsc@hirlam.org mailing list
- Specific technical issues -> who?
- Best way to keep good communication and feedback between HIRLAM and BSC?
- SBUs account
- Any other issue?


Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación
EXCELENCIA
SEVERO
OCHOA
Thank you
xavier.yepes@bsc.es